UPME: An Unsupervised Peer Review Framework for Multimodal Large Language Model Evaluation

Supplementary Material

7. Model Selection

We select models as follows:

GPT-4o [35] A versatile multimodal model by OpenAI, handling text, image, and audio inputs. It excels in vision and language tasks with enhanced processing speed. Known for strong real-time performance in audio and vision, GPT-4o is ideal for a variety of applications, including multilingual tasks.

GPT-4o mini [33] A smaller, cost-effective version of GPT-4o, optimized for handling text and images, with plans for audio support. It is designed for high-volume, real-time applications like chatbots and coding tasks, offering strong performance at a lower cost.

Gemini-1.5-Pro [45] Developed by Google DeepMind, this model uses Mixture-of-Experts architecture to optimize performance. It supports up to 1 million tokens and excels in translation, coding, and multimodal tasks. It is ideal for enterprise use due to its cost-efficiency and scalability.

Claude-3.5-Sonnet [3] From Anthropic, this model is optimized for reasoning, coding, and multimodal tasks. It excels in complex problem-solving and visual understanding, making it useful for customer support and detailed code-generation tasks.

Claude-3-Haiku [4] Developed by Anthropic, Claude 3.5 Haiku is a high-speed language model optimized for rapid response and advanced reasoning. With a 200K token context window and a maximum output of 4,096 tokens, it efficiently handles large datasets. Its affordability and speed make it ideal for applications requiring quick, concise responses, such as interactive chatbots and real-time data analysis.

Llama-3.2-11B-Vision-instruct [29] a multimodal large language model from Meta with 11 billion parameters, designed to handle both text and image inputs. It excels in tasks such as image captioning, visual question answering, and interpreting complex visual data. This model is particularly effective for industries like healthcare and retail, where real-time visual and textual analysis is key.

8. Extended Experiment

8.1. Pre-experiment for Different Weights.

We have conducted a preliminary experiment to investigate the relationship between the confidence weights and scores of models, as a substantiation of our methodology. Below is a detailed description of the experiment and its findings, which align with the methodology discussed in the paper.

Method	MMstar		ScienceQA	
	Pearson	**Spearman**	**Pearson**	**Spearman**
Reverse Weight	0.607	0.486	0.334	0.257
Uniform Weight	0.725	0.771	0.463	0.686
Consistent Weight	0.807	0.829	0.760	0.771
UPME	**0.944**	**0.972**	**0.814**	**0.886**

Table 4. Performance comparison of Consistent, Uniform and Reverse weight.

To begin, we designed a toy experiment to examine the role of confidence weights w. Based on the scores on manually designed benchmarks, we can obtain a model score ranking list, then designing weight configurations that are either consistent or reverse to this score ranking. Specifically, we constructed three weighting methods: Reverse Weight ($w = [0, 0.1, \ldots, 1]$), Uniform Weight ($w = [1, 1, \ldots, 1]$), and Consistent Weight ($w = [1, 0.9, \ldots, 0]$). Using these manually constructed weight configurations, we calculated the response score G_j for each model based on the predefined Equation 6 in Section 3 and obtained the score list $\hat{G}$ for all models. We then measured the alignment between the obtained ranking $\hat{G}$ and the human-annotated score list G using predefined metrics.

The results as summarized in Table 4, demonstrate that the Consistent Weight configuration achieves the highest correlation values, while the Reverse Weight configuration consistently yields the poorest results. These findings validate the proposed consistency assumption: assigning higher weights to models with stronger capabilities leads to better alignment between the model score list and the human-annotated one. In essence, responses recognized more favorably by other "reviewers" (models) tend to originate from higher-level models. This reinforces the idea that high-capability MLLMs evaluate others' responses more accurately and achieve higher scores.

We formalize this observation as the consistency assumption, which states that: 1. High-level LLMs exhibit greater confidence and accuracy when evaluating responses compared to lower-level ones. 2. A model's ability and its associated score are generally consistent.

Building on this preliminary finding, we devised an optimization framework aimed at maximizing the consistency between each model's capability (w) and its response score (G), constrained by our proposed methodology.

8.2. More Datasets

We have experimented on MMVet and the results in Table 5 show that UPME maintains its superior performance.

Models	MMstar		ScienceQA		MMVet	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
Peer Review	0.725	0.771	0.463	0.686	0.688	0.752
Majority Vote	0.757	0.757	0.509	0.524	0.732	0.643
Rating Vote	0.795	0.743	0.623	0.629	0.739	0.743
PRD [22]	0.838	0.864	0.692	0.636	0.794	0.814
UPME	0.944	0.972	0.814	0.886	0.914	0.928

Table 5. Comparison with recent methods.

8.3. Hyperparameter

The weights γ_1, γ_2, and γ_3 in Equation 9 were initialized as 0.4, 0.4, and 0.2, respectively, reflecting a balanced emphasis on response correctness and visual understanding while slightly de-emphasizing image-text correlation. This choice is intuited that correctness and reasoning typically have a larger impact on multimodal evaluation tasks.

γ_1	γ_2	γ_3	Pearson	Spearman
0.4	0.4	0.2	0.9415	0.9441
0.3	0.3	0.4	0.9397	0.8581
0.5	0.3	0.2	0.9306	0.7174
0.3	0.5	0.2	0.9365	0.8857

Table 6. hyperparameter in Scoring criteria.

To validate the optimality of this combination, we conducted experiments with different hyperparameter configurations. The results for four representative settings are summarized in Table Table 6. The proposed setting achieves the highest Pearson and Spearman correlations, indicating its effectiveness in aligning with human evaluations.

Notably, our experiments also show that the framework is relatively insensitive to small variations in these weights within the range [0.2,0.5], demonstrating its robustness.

Task-Specific Flexibility: The UPME framework supports task-specific flexibility. For instance, users may adjust γ_3 to prioritize image-text correlation in tasks requiring strong alignment between modalities or increase γ_2 for tasks demanding advanced reasoning capabilities, which allows the framework to cater to diverse evaluation needs.

Future Directions: While manual tuning of hyperparameters has proven effective, we agree that automating this process would further enhance the framework's generality and ease of use. We are actively exploring automated methods, such as validation-based optimization techniques or reinforcement learning approaches, to dynamically determine these weights based on task characteristics.

8.4. Reliability of Judge Correctness

Advantages of UPME's Question Generation: In the original peer review mechanism, the judge model might encounter questions that it cannot answer accurately, leading to unreliable evaluations. In contrast, the UPME framework enables the judge model to generate questions autonomously, ensuring that these questions fall within its capability. This significantly enhances the reliability of the judge model in assessing the correctness of responses from the evaluated models.

Empirical Evidence of Reliability: As shown in Table 7, UPME demonstrates a substantial improvement in both accuracy and human agreement compared to the original peer review mechanism.

Method	Dataset	Accuracy (%)	Human (%) Alignment
Peer Review	MMStar	64.2	71.1
	ScienceQA	60.3	68.2
UPME	MMStar	87.8	95.9
	ScienceQA	79.6	87.4

Table 7. $Judge_{Correct}$ reliability.

9. More Information about UPME

9.1. The Cost of UPME

UPME significantly reduces both time and financial costs: **Time Costs:** Creating VQAs manually requires deep understanding and may take several minutes to hours per task. AI tools significantly reduce this time, with labeling efficiency improved by up to 100 times [39]. UPME further accelerates evaluation, processing dozens of images per second. **Financial Costs:** Human annotations cost 1–5 per image depending on complexity, while UPME reduces this to approximately 1/7 of the manual cost.

Baseline methods like PeerReview and Majority Vote require extensive human-labeled data, significantly increasing time and costs. UPME eliminates manual annotation, offering efficient evaluations.

Method	Time / img	Finance / img
Human Annotation	3∼10 min	$ 2∼7
Majority Vote Rating Vote	2∼8 min	$ 1∼5
UPME Framework	1.5 s	$ 0.15

Table 8. Comparison of Time and Financial Costs

9.2. Image-Text Correlation

To compute the Image-Text Correlation score S_{Clip}, we employ the CLIP model [37], which measures the cosine similarity between image embeddings and text embeddings. For textual responses $A_i^{j,r}$ exceeding CLIP's maximum token input limit of 77 tokens, we implement a segmentation strategy that ensures each segment contains no more than the limit, preserving the context across the text.

Specifically, if the number of tokens n in a response exceeds 77, we calculate the starting indices for each segment by dividing the range from 0 to $n - 77$ into five equal intervals. These numbers serve as the starting points for each segment. Each segment then extends for 77 tokens from its starting index, ensuring full coverage of the original text with some overlap between consecutive segments. This overlapping is crucial as it helps preserve the continuity and context of the text, which might otherwise be lost if the segments were disjointed. The segments are then processed alongside the image through the CLIP model to generate embeddings, and cosine similarities between each text segment and the image are calculated. We derive the average of these similarity scores to evaluate the text-image alignment.

A notable feature of the segmentation strategy addresses potential verbosity bias by penalizing segments containing irrelevant or poorly aligned content. By computing the average cosine similarity across all segments, the approach inherently discounts segments that introduce irrelevant or poorly aligned content, reducing the score for long but less relevant responses. This mechanism effectively counteracts the verbosity bias of MLLM-as-a-judge.

9.3. Algorithm of UPME

Algorithm 1 Algorithm of UPME

Input: MLLM Pool $\mathcal{M}$, Image pool $\mathcal{I}$, Epochs T
Output: Model Scores $\hat{G}$, Weights w

Initialize w and G for models in M
// Dynamic Update $\hat{G}$ and w
for each iteration $t = 1$ to T **do**
    Randomly select $M_r, M_j, M_k \in M$
    Generate $Q_i^r, A_i^{j,r}, A_i^{k,r}$
    Calculate $S_{VL}(A_i^{j,r}, A_i^{k,r}, Q_i^r, I_i | M_r)$
    Update scores using EMA:
    $G[M_j] \leftarrow (1-\alpha)G[M_j] + \alpha S_{VL}$
    $w \leftarrow \text{optimize\_weights}(G)$
end for

9.4. Human Preference Alignment

The annotation work for human preferences alignment was carried out by five human experts with professional English proficiency, taking them a total of 170 hours. The labeling

Figure 8. Screenshot of human annotation.

screenshot is shown in Figure 8. The guidelines for human annotation are shown in Figure 9. Each data is associated with an image, a review model, and two candidate models, and it requires the completion of the following two annotation tasks: 1) Without knowledge of the review model's judgment, the annotator provides their own choice. 2) After being informed of the review model's judgment, the annotator indicates whether they agree with the decision.

These two tasks are assigned to different individuals for the same image, meaning that the same annotator does not perform both tasks for the same image. Each task is annotated by two annotators. When the results of the two annotators are consistent, the image's human preference annotation is obtained. If the results are inconsistent, up to five annotators will vote, and a majority vote determines the final annotation result for the data. Statistical analysis shows that such cases requiring voting account for only 2.17% of the final annotation results.

The experimental results on Table 3 indicate that the baseline method exhibited relatively low human agreement and model consistency rates, suggesting that the Peer Review mechanism under an unsupervised setting without weight optimization struggles to align with human preferences. In contrast, UPME demonstrated significant improvements by incorporating metrics such as Correctness, Visual Understanding, and Clip Relevance. On the MMstar dataset, UPME achieved an agreement rate of 95.9% and a consistency rate of 89.8%, showing that the optimized scoring criteria significantly enhance the accuracy of evaluation outputs and alignment with human preferences. By capturing key multimodal understanding metrics without relying on manual labeling, UPME effectively achieved high consistency with human annotations, highlighting its substantial advantages in improving response consistency and accuracy under an unsupervised framework.

Human annotation guideline

[Task Overview]

You are a human expert tasked with annotating the data assigned to you. You need to evaluate the responses of two candidate models to a given image and question, make your judgment, and assess whether the review model's judgment is correct. Each annotation involves assessing data instances that include an image, the responses from two candidate models, and a judgment from the review model.

For the given data, you will perform one of two tasks and focus your assessment on one of two aspects. Please be aware of which task the data belongs to and which aspect of the candidate models' responses you are evaluating.

[Two Annotation Tasks]

Task 1: Independent Choice Without Review Model Judgment
- You should independently evaluate and select your preferred response between the two candidate models based on their response to the image-related question.
- No information about the review model's judgment is provided during this step.

Task 2: Agreement with Review Model Judgment
- You are informed of the review model's judgment and asked to decide whether you agree or disagree with it.

Note: Tasks 1 and 2 must be conducted by different individuals for the same image to prevent cognitive bias.

[Two Aspects to Evaluate]

When evaluating the responses from the two candidate models, you need to focus on one of the following two aspects:

1) Correctness
- Your evaluation should be strictly objective, focusing only on which response is correct. Please proceed as follows:
- If only one model provides a correct answer, identify the correct model.
- If both answers are correct or both are incorrect, output 'C' to indicate a tie.

2) Visual Understanding and Reasoning
- Focus exclusively on the depth of visual understanding and the quality of reasoning in each response. Do not evaluate based on correctness. Here are the Evaluation Criteria:
- Captioning: Evaluate the ability to generate precise descriptions of image elements.
- Reasoning: Assess logical consistency and coherence in explanations and conclusions.
- Grounding: Evaluate accurate object localization within the image.
- Relationship: Assess the understanding of relationships and interactions between subjects in the image.

Figure 9. Human annotation guideline.

10. Prompt Template

Question generation prompt for judge model

You are a judge model tasked with evaluating the visual capabilities of two other models.

Based on the provided image input, generate a question that is directly related to the content of the image.

Please respond with only the question and no additional content.

Judge prompt for judge model focusing on visual understanding and reasoning

[System]
You are a judge model tasked with evaluating responses from two assistants to a question about an image.

Each assistant has provided an answer based on their analysis of the image.

Evaluation Criteria:
- Captioning: Evaluate the ability to generate precise descriptions of image elements.
- Reasoning: Assess logical consistency and coherence in explanations and conclusions.
- Grounding: Evaluate accurate object localization within the image.
- Relationship: Assess the understanding of relationships and interactions between subjects in the image.
- Focus exclusively on the depth of visual understanding and the quality of reasoning (as described above) in each response. Do not evaluate based on correctness.

Evaluation Format:
- Compare the two responses impartially. Ignore the order of presentation and the length of the responses. Do not favor any specific assistant based on their name.

- Conclude your evaluation by using the following format:
- [[A]] if assistant A's response demonstrates better visual understanding and reasoning,
- [[B]] if assistant B's response demonstrates better visual understanding and reasoning,
- [[C]] if it is a tie.

[User Question]
{question}

[The Start of Assistant A's Response]
{Answer_a}
[The End of Assistant A's Response]

[The Start of Assistant B's Response]
{Answer_b}
[The End of Assistant A's Response]

[Task]
Based on the image, question, and two responses provided, and following the criteria above, determine which assistant provided a better answer focusing solely on visual understanding and reasoning. Use the specified format for your final verdict.

Judge prompt for judge model focusing on correctness

[System]

You are a judge model tasked with evaluating responses from two assistants to a question about an image. Each assistant has provided an answer based on their interpretation of the image.

Your evaluation is strictly objective, focusing only on which response is correct. Please proceed as follows:

1. Assess if Assistant A's response is correct.
2. Assess if Assistant B's response is correct.
3. Compare the correctness of both responses:
- If only one assistant provides a correct answer, output "[[A]]" if Assistant A is correct, or "[[B]]" if Assistant B is correct.
- If both answers are correct or both are incorrect, output "[[C]]" to indicate a tie.

Avoid considering subjective factors such as response quality, detail, or reasoning process. Base your decision solely on the correctness of the answers.

[User Question]
{question}

[The Start of Assistant A's Response]
{Answer_a}
[The End of Assistant A's Response]

[The Start of Assistant B's Response]
Answer_b
[The End of Assistant B's Response]

[Task]

Based solely on the correctness of the two responses, determine which assistant answered the question accurately. Use the specified format for your final verdict.

Answer generation prompt for candidate model

[System]
Please act as an image-understanding expert to solve the problem based on the provided image.

First, analyze the provided image in detail, focusing on its overall theme and key elements.

Then, outline your reasoning process step by step, considering how each detail contributes to your understanding of the image.

Finally, provide a clear and accurate answer to the user's question based on your analysis. Let's think step by step.

[User Question]
{question}

Once you've completed your reasoning, pick one choice from the options. Output the final answer in the format: "[[X]]" where X is the selected option.